EVALUATION OF A COMPUTABLE PHENOTYPE FOR IDIOPATHIC PULMONARY FIBROSIS

A Thesis in
Public Health Sciences
by
Anne E.F. Dimmock

© 2016 Anne E.F. Dimmock

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2016
The thesis of Anne E.F. Dimmock was reviewed and approved* by the following:

Cynthia H. Chuang  
Professor of Medicine and Public Health Sciences  
Thesis Advisor

Wenke Hwang  
Associate Professor of Public Health Sciences

Herbert Y. Reynolds  
Emeritus Professor of Medicine

Junjia Zhu  
Assistant Professor of Public Health Sciences

Kristen H. Kjerulff  
Professor, Department of Public Health Sciences  
Director, Master of Science in Public Health Sciences Program

*Signatures are on file in the Graduate School.
Abstract

Introduction: The electronic medical record (EMR) is a common source of data for clinical research. However, the quality of EMR-based research depends on the validity of computable phenotypes, defined as computerized queries to an EMR system or clinical data repository to identify individuals with a clinical condition. Computable phenotyping is especially of interest with rare diseases, such as idiopathic pulmonary fibrosis (IPF). Our objective was to evaluate a computable phenotype for the identification of patients with idiopathic pulmonary fibrosis.

Methods: A computable phenotype was adapted from previously published algorithms, which identified patients in our health system if they had at least one ICD-9 code of 516.3 (idiopathic interstitial pneumonia) or 516.31 (idiopathic pulmonary fibrosis) in the past three years associated with inpatient or outpatient visits, excluding emergency department and lab encounters. Patients were excluded if an ICD-9 code for connective tissue disorders was present. Gold standard diagnosis was determined by chart review by at least two reviewers using a standardized procedure. In cases of disagreement, charts were reviewed collectively until agreement was reached.

Results: The computable phenotype identified 157 individuals with IPF. After gold standard chart review, 74 patients (47%) were classified as having IPF based on a narrow definition and 89 (57%) were classified as having IPF based on a broad definition.

Conclusion: Our computable phenotype yielded a positive predictive value of 47% using a narrow definition of IPF. True positives were older (p=0.0003) than false positives. Studies depending on computable phenotypes for identifying patients with IPF will be limited by false positives.
TABLE OF CONTENTS

List of Tables ..............................................................................................................v
List of Abbreviations ..................................................................................................vi
Acknowledgements ..................................................................................................... vii
Introduction ...............................................................................................................1
   Idiopathic Pulmonary Fibrosis ..............................................................................1
   The Electronic Health Record ............................................................................3
   Computable Phenotypes .....................................................................................3
   PaTH Network .....................................................................................................4
   Specific Aims ......................................................................................................5
Methods .....................................................................................................................6
   Data Source .........................................................................................................6
   Development of Our Computable Phenotype ....................................................7
   Application of the Computable Phenotype to the Electronic Health Record .......8
   Chart Review Procedure ....................................................................................5
   Statistical Methods ............................................................................................5
Results ......................................................................................................................11
Discussion and Conclusions ...................................................................................13
Appendices ...............................................................................................................19
   Appendix A: Exclusionary ICD-9 Codes ............................................................19
   Appendix B: Standardized Chart Review Procedure ........................................20
   Appendix C: Diagnostic Algorithm Used for Chart Review ..............................20
References ...............................................................................................................21
List of Tables

Table 1: Data Sources Available in i2b2 ................................................................. 5
Table 2: Consensus Reviewer (Gold Standard) Diagnosis from Chart Review ..................11
Table 3: Comparison of Demographics of True and False Positives ..............................12
List of Abbreviations

ATS: American Thoracic Society
BAL: Bronchoalveolar Lavage
CDRN: Clinical Data Research Network
COPD: Chronic Obstructive Pulmonary Disease
CPFE: Combined Pulmonary Fibrosis and Emphysema
CT: Computed Tomography
DIP: Diffuse Interstitial Pneumonia
EHR: Electronic Health Record
FPF: Familial Pulmonary Fibrosis
HRCT: High Resolution Computed Tomography
i2b2: Integrating Informatics from the Bench to the Bedside
ICD: International Classification of Diseases
ILD: Interstitial Lung Disease
IPF: Idiopathic Pulmonary Fibrosis
NLP: Natural Language Processing
NSIP: Non-specific Interstitial Pneumonia
PCORI: Patient Centered Outcomes Research Institute
PFNOC: Pulmonary Fibrosis, Not Otherwise Classifiable
PNPRC: PaTH Network Protocol Review Committee
PPV: Positive Predictive Value
SLB: Surgical Lung Biopsy
UIP: Usual Interstitial Pneumonia
Acknowledgements

Thank you to my family and friends for your support and encouragement.

This project would not have been possible without David Meck, who provided assistance with i2b2. I would also like to thank the other members of the PaTH team in Hershey: Jody McCullough, Bobbie Johannes, Bari Dzomba, and Sanjoy Bhattacharjee.
**Introduction**

**Idiopathic Pulmonary Fibrosis**

Idiopathic pulmonary fibrosis (IPF) is a rare disease of the lungs, which involves progressive scarring and is often fatal within 3 to 5 years. IPF is characterized by dyspnea and cough in its early stages. Patients often require supplemental oxygen as the disease progresses and they become increasingly hypoxemic.

The definition and diagnostic criteria for pulmonary fibrosis have changed over time. Pulmonary fibrosis was first described as an entity in the 1940s, and was referred to at the time as Hamman-Rich syndrome(1). The first case report of Hamman-Rich Syndrome was published in 1952, and was described as an acute diffuse interstitial fibrosis of the lungs.(2) Several decades later, a publication by Carrington et al set out to describe the natural history of two types of pulmonary fibrosis: usual interstitial pneumonia (UIP) and diffuse interstitial pneumonia (DIP).(3)

Before there were high-resolution scans of the lungs, the diagnosis of pulmonary fibrosis was often made using a combination of a chest x-ray images and lung biopsy.(4, 5) Additional analysis using bronchoalveolar lavage (BAL) was also common.(6, 7) However, the diagnostic approach has evolved over time as high resolution CT scans have become more common. This has made lung biopsies less common and not often needed in diagnosis.

As it is currently understood, IPF is a chronic and progressive lung disease characterized by lung scarring. The diagnosis requires exclusion of other known causes of pulmonary fibrosis, as well as a radiographic and/or histologic pattern of usual interstitial
pneumonia (UIP) suggested on high resolution computed tomography (HRCT) and/or on histology from surgical lung biopsy (SLB), respectively.

Just as the diagnosis of IPF has changed significantly, treatment of IPF has undergone advances in recent years(8); most recently there has been the addition of the first two FDA-approved drugs to treat the disease, which were both approved in 2014.(9, 10) The two medications, Esbriet® (pirfenidone, Roche-Genentech) and OFEV® (nintedanib, Boehringer Ingelheim), each demonstrated a reduction in the decline in forced vital capacity at 52 weeks, when compared with placebo.

The availability of drug therapies to treat (albeit not cure) this disease has led to increasing interest in earlier diagnosis. In addition, companies are understandably interested in the epidemiology of IPF in order to quantify the potential market for new therapies. However, the only “cure” for pulmonary fibrosis is lung transplantation. Absent this, median survival is approximately 3-5 years. A recently published mortality prediction model bases prognosis on a patient’s gender, age, and lung function.(11)

Pulmonary fibrosis is considered a rare disease, in that fewer than 200,000 people in the US are thought to have the disease. This follows the definition for a rare disease laid out in the Orphan Drug Act of 1983, which has become the standard definition in the United States.(12) There are over 7000 rare diseases known in the US(13), making rare diseases a topic of great interest.

Previous studies of the epidemiology of IPF have relied primarily on claims data. As this is a disease of older individuals, Medicare claims data have been used to estimate the epidemiology of IPF. A 2006 study using Medicare data estimated the prevalence of IPF to be 14-43/100,000 persons.(14) A publication using clinical data from individuals living in
Olmstead County, Minnesota estimated the prevalence to be 27.9 cases/100,000 persons. (15) A publication using a private insurance claims database estimated the prevalence of IPF in the US to be 125.2/100,000 persons, among patients ages 50 and older. (16)

The apparent increase in incidence and prevalence of IPF may be due to reporting and recognition bias, and not due to an increased burden of disease in the population. (17) However, the aging US population may contribute to an increased incidence of IPF by 2030. (15)

The Electronic Health Record (EHR)

An electronic health record, or EHR, is a digital repository of a patient’s health information. More broadly, the electronic health record system refers to both the digital infrastructure (e.g., software, network) and the data (i.e., patient information) that together form a health system's library of patient data. EHR use has continued to increase in the United States, with over 75% of physicians reportedly using an EHR in 2013. In 2001, only 18% of physicians reported using an EHR. (18) This drastic growth in the use of EHRs is due, in part, to legislation that established incentive programs to encourage hospitals and other healthcare providers to adopt EHRs for use in routine care. (19)

In addition to its primary function as a tool for routine clinical care, the electronic health record has also been expanded as a tool for clinical research (20) in some instances. Finally, the electronic health record is an important source of data for public health surveillance, potentially improving the timeliness and accuracy of data reporting. (21)

Computable Phenotypes

A computable phenotype is defined as “a clinical condition or characteristic that can be ascertained via a computerized query to an EHR system or clinical data repository using
a defined set of data elements and logical expressions. These queries can identify patients with a particular condition... and can be used to support a variety of purposes and data needs for observational and interventional research”(22).

One strength of a computable phenotype is that it does not require time-consuming chart review by a clinician or other trained reviewer in order to identify cases. However, as with all methods in health services research, computable phenotypes are limited by the quality of available data. Computable phenotypes are of particular interest to the rare disease community, because of their potential to better identify patients for both observational studies and clinical trials. Computable phenotype definitions using electronic health record data (EHR) have also been published on more common conditions, such as diabetes mellitus(23), chronic obstructive pulmonary disease (COPD)(24), and low back pain(25). Other computable phenotypes have been tested using a combination of EHR data and genomics data, through the eMERGE network.(26) eMERGE has utilized registries with combined clinical data and biospecimens, using computable phenotypes to select individuals for genetic sequencing based on characteristics of interest.

PaTH Network

The PaTH Network is one of 13 Clinical Data Research Networks (CDRNs) funded by PCORI, the Patient Centered Outcomes Research Institute. The goal of these CDRNs is to link electronic health data with patient reported outcomes. The PaTH Network is comprised of four academic medical centers in the mid-Atlantic region: University of Pittsburgh, Penn State University, Temple University, and Johns Hopkins University. Data from each institution’s electronic health record flows through i2b2 and SHRINE into a
single data repository housed at the University of Pittsburgh(27). PaTH chose to study three conditions initially: atrial fibrillation, obesity, and idiopathic pulmonary fibrosis.

Specific Aims

The purpose of this study was to determine the positive predictive value of a computable phenotype for idiopathic pulmonary fibrosis using data from the electronic health record. This was achieved through four specific aims:

1) To develop an algorithm to identify IPF patients;
2) To apply the algorithm to an electronic health record,
3) To assign a gold standard diagnosis to each patient based on a chart review, and
4) To evaluate the algorithm using two IPF classifications.
Methods

This research was reviewed by the PaTH Network Protocol Review Committee (PNPRC)\(^1\) approved by the Institutional Review Board at Johns Hopkins. The protocol was also reviewed and approved by the Institutional Review Board at Penn State Milton S. Hershey Medical Center.

Data Source

The i2b2 application allows researchers to execute queries against a single framework, which is comprised of a multitude of data types that have been combined from numerous and varying sources. Data from the hospital's health record system, including billing and registration data (Eclypsis), laboratory results and provider-based health records (Cerner®), are all combined in the i2b2 system. In addition to data from the electronic health record, i2b2 also allows users to import data from other sources, such as REDCap for patient-reported outcomes, into a project database (Table 1).(28) Data from the source systems are loaded into an intermediary framework, called the Operational Data Store (ODS).

Table 1: Data Sources Available in i2b2

<table>
<thead>
<tr>
<th>Source</th>
<th>Data Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerner</td>
<td>Patient</td>
</tr>
<tr>
<td></td>
<td>Visit</td>
</tr>
<tr>
<td></td>
<td>Lab Results</td>
</tr>
<tr>
<td></td>
<td>Medications</td>
</tr>
<tr>
<td>Eclipsys</td>
<td>ICD</td>
</tr>
<tr>
<td></td>
<td>DRG</td>
</tr>
<tr>
<td>DSAR/Signature</td>
<td>Procedure Code</td>
</tr>
<tr>
<td>REDCap</td>
<td>Survey Data</td>
</tr>
<tr>
<td>TNM</td>
<td>Cancer Registry</td>
</tr>
</tbody>
</table>

\(^1\) The PNPRC was established to ensure input from each institution prior to review by the central IRB at Johns Hopkins University. The PNPRC has representation from at least one local IRB official and one patient representative from each PaTH institution.
Patients with an electronic medical record (Cerner®) at Hershey Medical Center were included in the source population. This population includes all patients who have had any clinical encounter between January 1, 2011 and December 31, 2015. All visits and data for patients who meet those criteria are included in the dataset, including visits that occurred prior to January 1, 2011.

**Development of Our Computable Phenotype**

The PaTH IPF Working Group\(^2\) agreed upon a PaTH IPF computable phenotype after reviewing the literature for previously published algorithms\(^{14, 29}\) and consensus of expert opinion. Inclusion criteria were patients in our health system if they had at least one ICD-9 code of 516.3 (idiopathic interstitial pneumonia) or 516.31 (idiopathic pulmonary fibrosis) in the past three years, defined as 12/1/2013 through 11/30/2016, associated with inpatient or outpatient visits, excluding emergency department and lab encounters. Patients were excluded if an ICD-9 code for connective tissue disorders was present, using ICD-9 codes based on a list by Raghu et al.\(^{14}\) Connective tissue disorders, also known as collagen vascular diseases, are systemic autoimmune diseases that often include respiratory manifestations in the form of interstitial lung disease.\(^{30}\) The final list of exclusionary ICD-9 codes is included in Appendix A. Some modifications were made to the list of exclusionary ICD-9 codes after discussion among IPF experts in the group. These included the addition of: 1) ICD-9 code 279.49, Antisynthetase syndromes, and 2) ICD-9 code 710.9, undifferentiated connective tissue disease. The group considered adding ICD-9

---

\(^2\) The PaTH IPF Working Group is comprised of IPF experts (clinician-researchers) and IPF patient partners. The Working Group meets weekly to review protocols and harmonize terminology and process for IPF-related research in the PaTH Network.
code 443.0, Raynaud’s phenomena, as an exclusionary diagnosis, however after discussion elected not to include this, since Raynaud’s is prevalent and its presence on the list may unnecessarily exclude people.

Application of the Computable Phenotype to the Electronic Health Record

Patients with an electronic medical record (Cerner®) at Hershey Medical Center were included in the source population. This population includes all patients who have had any clinical encounter between January 1, 2011 and December 31, 2015. An i2b2 interface was used to query the source population for patients who met the definition of idiopathic pulmonary fibrosis as outlined in the computable phenotype.

In summary, the computable phenotype is employed in the process of loading patient data into i2b2. The query representing the computable phenotype definition was executed against ODS, and the patients identified from this query were loaded into i2b2, with an identifying flag for the IPF cohort. EHR data were loaded into i2b2 by mapping the data to standard vocabularies (the “Extract, Transform, Load” process).

Chart Review Procedure

Gold standard diagnosis was determined by independent chart review by at least two reviewers using a standardized procedure (see Appendix B) that was developed based on the international consensus statement Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management. Figure 3 in this document provides a diagnostic algorithm for IPF, from which the chart review procedure was derived, as shown in Appendix C.(31) The standardized chart review procedure included review of pulmonary outpatient notes, radiology interpretation of chest CT scan(s), and pathology reports from surgical lung biopsies, when available. Chest CT scans were recorded for presence/absence
of usual interstitial pneumonia (UIP) pattern (per radiology report) and presence/absence of honeycombing. Pathology reports were recorded for presence/absence of UIP pattern. History of an occupational or environmental exposure and indication of connective tissue serology/rheumatology diagnosis was also recorded. Based on the above information, each reviewer categorized each patient as definitely having idiopathic pulmonary fibrosis (IPF), familial pulmonary fibrosis (FPF), combined pulmonary fibrosis and emphysema (CPFE), other pulmonary fibrosis, e.g. NSIP, radiation-induced ILD, occupational or environmental ILD connective tissue-associated ILD, granulomatous disease, or non-ILD. IPF, FPF, and CPFE were considered to be true diagnoses for the broad definition. Only IPF was considered true IPF for the narrow definition. In cases where there was insufficient information in the EHR, the diagnosis was listed as unknown. The reviewers convened to review cases where there was disagreement. During this reconciliation process, charts were reviewed collectively until agreement was reached.

Each patient identified by the PaTH IPF computable phenotype was subsequently characterized as a true positive or a false positive, based on the gold standard chart review. Positive predictive value (PPV) was calculated to describe the accuracy of the computable phenotype against the consensus diagnosis based on chart review (gold standard).

**Statistical Methods**

The point estimate of positive predictive values was reported and its 95% confidence interval was generated. Comparisons of true positive and false positive findings were done using either a Fisher's exact test for categorical factors (e.g., race, gender) or a two-sample student's t-test for continuous variables (e.g., age).
Statistical comparison of true positives and false positives used two different definitions for a gold standard: the narrow definition and the broad definition. The narrow definition included only cases with a consensus diagnosis of “IPF”. The broad definition included cases with a consensus diagnosis of “IPF”, “FPF”, or “CPFE”.

All statistical tests were done using SAS® version 9.4. All statistical tests are two-sided and the significance level used was 0.05.
Results

One hundred fifty seven patients were identified as having IPF by the computable phenotype. As shown in Table 2, after chart review, 74 patients (47%) were classified as having true IPF using the narrow definition. Another 40 (25.5%) were classified as having another form of pulmonary fibrosis, e.g., familial pulmonary fibrosis (6, 3.82%), combined pulmonary fibrosis and emphysema (9, 5.73%), non-specific interstitial pneumonia (8, 5.1%), or pulmonary fibrosis not otherwise classifiable (17, 10.83%).

The computable phenotype resulted in a positive predictive value of 47% (74/157) using the narrow definition. Using a more inclusive definition of IPF for the gold standard comparison, the positive predictive value increased to 57% (89/157). The additional true positives came from cases with a gold standard diagnosis of familial pulmonary fibrosis (n=6) or combined pulmonary fibrosis and emphysema (n=9).

Table 2: Consensus Reviewer (Gold Standard) Diagnosis from Chart Review

<table>
<thead>
<tr>
<th>Consensus Reviewer (Gold Standard) Diagnosis</th>
<th>Patients (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiopathic Pulmonary Fibrosis (IPF)</td>
<td>74 (47.13%)</td>
</tr>
<tr>
<td>Pulmonary Fibrosis, not otherwise classifiable</td>
<td>17 (10.83%)</td>
</tr>
<tr>
<td>Combined Pulmonary Fibrosis and Emphysema</td>
<td>9 (5.73%)</td>
</tr>
<tr>
<td>Familial Pulmonary Fibrosis</td>
<td>6 (3.82%)</td>
</tr>
<tr>
<td>Granulomatous Diseases, e.g. Langerhans</td>
<td>3 (1.91%)</td>
</tr>
<tr>
<td>Connective Tissue-Associated ILD</td>
<td>6 (3.82%)</td>
</tr>
<tr>
<td>Occupational, Radiation or Drug-Induced ILD</td>
<td>6 (3.82%)</td>
</tr>
<tr>
<td>Non-ILD</td>
<td>6 (3.82%)</td>
</tr>
</tbody>
</table>
Based on the information in the electronic health record, 63% of identified patients were alive and 37% were deceased. Sixty percent of identified patients were male, with a mean age of 71.7 years. 89% of identified patients were white. Patients with IPF using the narrow definition were significantly older than those without IPF (p=0.0003). There was no significant difference in age when using the broad definition. Neither definition had a significant difference in gender (% male) or race (% white).

Table 3: Comparison of Demographics of True and False Positives

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Narrow Gold Standard Diagnosis</th>
<th>Broad Gold Standard Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPF</td>
<td>Not IPF</td>
</tr>
<tr>
<td>% Male</td>
<td>60%</td>
<td>62%</td>
<td>60%</td>
</tr>
<tr>
<td>% White</td>
<td>89%</td>
<td>87%</td>
<td>91%</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>72 ± 13.8</td>
<td>75 ± 9.9</td>
<td>67 ± 16.6</td>
</tr>
</tbody>
</table>
Discussion

Our computable phenotype algorithm identified 157 individuals in our electronic health record (EHR) with IPF. After chart review to determine a consensus diagnosis, the algorithm had a positive predictive value of 47% when using the narrow gold standard definition, and 57% when using the broad gold standard definition. Using the narrow definition, true positives were significantly older (p=0.0003) than false positives (mean age of 75 and 67 years, respectively). There was no significant difference between the groups when using the broad definition. The inclusion of the FPF and CPFE cases in the broad definition decreased the mean from 75 years (narrow definition) to 73 years (broad definition). This may reflect a difference in etiology among IPF, FPF, and CPFE,

Other studies using multiple case definitions for EHR-based identification of patients have shown differences in performance based on how broad or narrow the definition was. For example, in patients with another pulmonary disease, chronic obstructive pulmonary disease (COPD), multiple case definitions, or computable phenotypes, were evaluated. Their identified gold standard was the “clinical trial standard”, which was comprised of pulmonary function testing results, and/or a history of smoking and alpha-one antitrypsin deficiency. This gold standard was compared with two other definitions: an ICD-9-based definition, and a patient-reported physician diagnosis.(24) They concluded that a clinical trial definition of COPD may not reflect the broader population of COPD patients, and that different case definitions will yield different populations. Similarly, the patients identified by our computable phenotype differ from the population used in clinical trials for IPF. This is particularly relevant as there are approved
therapies for IPF which, though tested in a narrowly-defined group of individuals, are now being prescribed to a broader group of patients who have been diagnosed with IPF.

For 18% of cases with a consensus diagnosis of “unknown”, there was not enough information available in the electronic health record for a definitive diagnosis. This is higher than the 10% of unclassifiable interstitial lung disease cases reported in other cohorts. In their study, the most common reason for being unclassifiable was the lack of histopathology from a surgical lung biopsy. Cases with a consensus diagnosis of “pulmonary fibrosis, not otherwise classifiable”, which comprised an additional 11 of the patients identified by the computable phenotype, may have fallen into the category of unclassifiable interstitial lung disease, given additional clinical evaluation. However, they may also have eventually been diagnosed with one of the other disease entities listed.

There are two possible approaches to handling PFNOC and unknowns. One option would be to exclude them from the analysis. The reason to exclude them would be that in both case there is insufficient information to come up with a clinical consensus diagnosis (vs research/chart review consensus diagnosis). The reason to include them, however, is akin to an intention to treat analysis. In using the EHR for research, a principle we favor is to define the population using EHR criteria and not to further restrict the population using non-EHR methodology. If you artificially remove people from the denominator, then you artificially inflate the performance characteristics of the computable phenotype in the EHR (smaller denominator = higher %). Recognizing this tradeoff, patients with a consensus diagnosis of pulmonary fibrosis, not otherwise classifiable, or unknown were not excluded from the analysis since they were part of the source population. An epidemiological study
of idiopathic pulmonary fibrosis using this EHR cohort would, without gold standard chart review, include these patients in their population of affected individuals.

Time period of our chart review was same time period as the information used for the computable phenotype algorithm. However, when looking at the chart for clinical purposes, this time restriction would not apply. One possible reason for the insufficient information is the artificial constraint of a two-year window.

Although we used the term PFNOC, it is not equivalent to the “nonclassifiable fibrosis” categorization outlined in the ATS consensus statement. Nonclassifiable fibrosis, as defined in the ATS statement, requires a histopathologic interpretation from a surgical lung biopsy; it is a pattern of fibrosis that does not meet criteria for usual interstitial pneumonia pattern or another idiopathic interstitial pneumonias. Our term, PFNOC, is defined as radiographic evidence of pulmonary fibrosis, without the availability of histopathology. This then represents incomplete ascertainment.

The low number of false positives with a gold standard diagnosis of connective tissue-associated ILD (n=6, 3.8%) indicates that the list of exclusionary ICD-9 codes did a good job of excluding patients in this category. By comparison, an estimated 15% of patients who present with interstitial lung disease have an underlying connective tissue disease.

**Strengths and Limitations**

Strengths of this study include the use of an algorithm based on previously published epidemiology studies, which allows the results to be compared to other epidemiologic estimates based on the same case definition. The use of the i2b2 platform
allows this study to be more easily replicated at other institutions, since i2b2 is used by several clinical data research networks (CDRNs) funded by PCORI.

Chart review was performed on 100% of the patients identified by the algorithm, which contrasts with other studies in which only a sub-sample was reviewed. By using a health system as the source population, instead of a claims database, we included all individuals regardless of insurance enrollment status. Another strength of the chart review process was the use of an international consensus statement on IPF diagnosis as the basis for the standardized chart review, as well as the use of two independent reviewers and the reconciliation process used when there was disagreement between reviewers. By using both a broad and narrow definition for a true positive, the results of this study will be applicable regardless of which definition the field settles on in the future. The computable phenotype was evaluated and described using standard metrics (e.g., PPV).

A significant limitation of this study is that we were unable to identify the false negatives from the source population, because chart review of the source population was not feasible. As a result, we are unable to accurately estimate the true prevalence of IPF in our population. Use of a patient registry or other source may help identify these cases in future studies.

This was a single-center study, which limits the generalizability of the results. Additional analyses in other electronic medical records and/or in other health systems will be needed. Our computable phenotype was also limited by the data available. Addition of variables such as presence of honeycombing on HRCT, as identified by natural language processing, would likely improve the PPV substantially. However, although additional variables may improve specificity, they will likely decrease sensitivity.
Variability in EHR and coding practices may influence the performance of this computable phenotype in other health systems. As with any computable phenotype, the results are also limited by the quality of EHR data. Further, one of the limitations of the EHR as currently constructed is the lack of searchable access across systems (the problem of “outside records”). This limitation is overcome by time-intensive manual searching by healthcare providers. These “outside records” were not included in our chart review process, since they are not indexed in our health record. In the case of IPF specifically, patients are often referred to a tertiary care center for diagnosis; initial testing such as lab results, imaging and pathology therefore may not be available in the EHR.

Future Directions

With the implementation of ICD-10, additional studies will need to be performed using the new coding scheme for IPF to see if this has improved the positive predictive value, weakened it, or whether it has remained approximately the same. One possible area of inquiry would be to look at the patients identified as having IPF using ICD-9 codes, and following how their disease is coded using ICD-10. Similarly, patients identified using ICD-10 coding could undergo retrospective chart review in order to see how their previous encounters were coded in the ICD-9 system.

As mentioned previously, this computable phenotype will need to be validated at other sites, both in the PaTH Network and external to the PaTH Network. Finally, future work needs to be done to identify the false negatives, i.e., those patients who have a diagnosis of IPF but were not identified using the computable phenotype.

Additional analysis comparing presence/absence of an HRCT and/or SLB may further improve the performance of the computable phenotype algorithm. Additional
variables for consideration are a primary diagnosis of IPF (compared to a secondary diagnosis of IPF) and an inpatient hospitalization for respiratory illness. A recently-published prediction equation also incorporates other forms of diagnostic testing, such as blood markers of underlying connective tissue disease, in their model. (16)

Finally, the use of natural language processing (NLP) provides a mechanism for accessing text-based interpretations from radiology and pathology and organizing them for future data analysis. This has been done successfully in a computable phenotype for rheumatoid arthritis. Their algorithm had a PPV of 88% when using coding data alone (i.e., ICD-9), and improved to 94% when narrative data were incorporated using NLP. (34) When the algorithm was applied at other institutions a new training set was used in order to improve the performance of the analysis based on site-specific variations in practice. (35) NLP could be used to identify honeycombing on imaging studies, or to capture possible exposures (e.g., living with birds) that may have been listed in a narrative from a clinician.
### APPENDIX A: Exclusionary ICD-9 Codes

<table>
<thead>
<tr>
<th>ICD-9-CM Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>Sarcoidosis</td>
</tr>
<tr>
<td>237.7</td>
<td>Neurofibromatosis</td>
</tr>
<tr>
<td>272.7</td>
<td>Lipidoses</td>
</tr>
<tr>
<td>277.3</td>
<td>Amyloidosis</td>
</tr>
<tr>
<td>277.8</td>
<td>Other specified disorders of metabolism—includes eosinophilic granuloma</td>
</tr>
<tr>
<td>279.49</td>
<td>Antisynthetase syndromes*</td>
</tr>
<tr>
<td>446.21</td>
<td>Goodpasture’s syndrome</td>
</tr>
<tr>
<td>446.4</td>
<td>Wegener’s granulomatosis</td>
</tr>
<tr>
<td>495.x (includes 495.0-495.9)</td>
<td>Extrinsic allergic alveolitis</td>
</tr>
<tr>
<td>500</td>
<td>Coal worker’s pneumoconiosis</td>
</tr>
<tr>
<td>501</td>
<td>Asbestosis</td>
</tr>
<tr>
<td>502</td>
<td>Pneumoconiosis due to other silica or silicates</td>
</tr>
<tr>
<td>503</td>
<td>Pneumoconiosis due to other inorganic dust</td>
</tr>
<tr>
<td>504</td>
<td>Pneumoconiosis due to inhalation of other dust</td>
</tr>
<tr>
<td>505</td>
<td>Pneumoconiosis, unspecified</td>
</tr>
<tr>
<td>506.4</td>
<td>Chronic respiratory conditions due to fumes or vapors</td>
</tr>
<tr>
<td>508.1</td>
<td>Chronic and other pulmonary manifestations due to radiation</td>
</tr>
<tr>
<td>508.8</td>
<td>Respiratory conditions due to other specified external agents</td>
</tr>
<tr>
<td>516.0</td>
<td>Pulmonary alveolar proteinosis</td>
</tr>
<tr>
<td>516.1</td>
<td>Idiopathic pulmonary hemosiderosis</td>
</tr>
<tr>
<td>516.2</td>
<td>Pulmonary alveolar microlithiasis</td>
</tr>
<tr>
<td>516.8</td>
<td>Other specified alveolar and parietoalveolar pneumonopathies</td>
</tr>
<tr>
<td>516.9</td>
<td>Unspecified alveolar and parietoalveolar pneumonopathies</td>
</tr>
<tr>
<td>517.2</td>
<td>Lung involvement in systemic sclerosis</td>
</tr>
<tr>
<td>517.8</td>
<td>Lung involvement in other diseases classified elsewhere</td>
</tr>
<tr>
<td>518.3</td>
<td>Pulmonary eosinophilia</td>
</tr>
<tr>
<td>555.x (includes 555.0, 555.1, 555.2, 555.9)</td>
<td>Regional enteritis</td>
</tr>
<tr>
<td>710.0</td>
<td>Systemic lupus erythematosus</td>
</tr>
<tr>
<td>710.1</td>
<td>Systemic sclerosis</td>
</tr>
<tr>
<td>710.2</td>
<td>Sjogren’s disease</td>
</tr>
<tr>
<td>710.3</td>
<td>Dermatomyositis</td>
</tr>
<tr>
<td>710.4</td>
<td>Polymyositis</td>
</tr>
<tr>
<td>710.9</td>
<td>Undifferentiated connective tissue disease*</td>
</tr>
<tr>
<td>714.81</td>
<td>Rheumatoid lung</td>
</tr>
<tr>
<td>720.0</td>
<td>Ankylosing spondylitis</td>
</tr>
</tbody>
</table>
APPENDIX B: Standardized Chart Review Procedure

1. Access patient chart using medical record number
2. Open most recent clinic note from pulmonary provider (outpatient note or outpatient letter)
   a. Mention of “pulmonary fibrosis” as chief complaint or in assessment/plan? Y/N
   b. Assessed as “idiopathic”? Y/N
3. Open first/previous clinic note from pulmonary provider (outpatient note or outpatient letter)
   a. Evidence of connective tissue/autoimmune disease? Y/N
   b. Evidence of clinically significant environmental or work exposure? Y/N
4. Open most recent chest CT (studies/chest/CT)
   a. CT scan available? Y/N/Outside only
   b. UIP pattern? Y/N
   c. Honeycombing? Y/N
5. Check for SLB (path/surgical)
   a. SLB available? Y/N
   b. UIP pattern? Y/N
6. Assign Reviewer Diagnosis

APPENDIX C: Diagnostic Algorithm Used for Chart Review

Adapted from Raghu et al AJRCCM 2006. A * indicates a diagnosis code added by PaTH that was not part of the original list by Raghu et al.
References


